247—252

23547(8)

动 物 学 研 究 1996, 17 (3): 247—252 CN 53-1040/Q ISSN 0254-5853
Zoological Research

综 述

从 DNA 序列到物种树*

张亚平
(中国科学院昆明动物研究所细胞与分子进化开放研究实验室 650223)

Q111

关键词 DNA 序列, 基因树, 物种树 进化
Key words DNA Sequence, Gene tree, Species tree

DNA 不仅是主要的遗传物质，同时也是生物进化史的重要记录者。通过 DNA 序列分析研究生物的进化历程、确定物种间的进化关系具有许多的优越性。第 1, DNA 仅由 4 种基本结构单位（G、A、T、C）组成，其序列上的异同是明确无误的，因而易于分析。第 2, DNA 序列含有无比丰富的进化信息。有些物种的基因组中具有多于 10^{11} 个碱基对。第 3, DNA 序列相对易于获取。特别是随着近年来 PCR 技术的应用与推广以及人类基因组项目的实施、DNA 序列正以爆炸性的速度积累起来。正是由于上述这些原因，DNA 序列分析已成为生物系统与演化研究中最重要与最热门的工具之一，并取得了许多令人瞩目的结果（张亚平, 1995; Miyamoto 等,1987;Hillis 等,1990;Zhang 等, 1993)。

以 DNA 序列研究物种的进化关系，大致分两大步骤：1）根据研究的对象与目的，选择适当的基因或其他 DNA 区域，并测定目标 DNA 片段的序列。对于近缘物种的研究，应选用进化速度比较快的区域；对于远缘物种，则应选用相对保守的区域。2）通过 DNA 同源序列的比较，采用一定的系统重建途径与方法，确定基因系统树和物种系统树。

如何正确地分析 DNA 序列以从中获取进化信息？这方面的研究已取得长足的进展，但尚有许多未能解决的问题。本文拟对系统研究中如何比较分析 DNA 序列作一简单的介绍与探讨。

1 同源 DNA 序列的排序（Alignment）

对两个同源 DNA 序列的比较，首先需要确定他们从最近的共同祖先分离后，各序列中缺失／插入所发生的位置以及同源部分的对应关系。这个过程叫排序。对于蛋白质编码区域而言，由于蛋白质功能上的需要和三联体密码结构的限制，缺失／插入很少发生或发生后很易被选择淘汰。因此，一般比较容易排序。而在非编码区域内，缺失／插入发生的

---

* 本工作得到中国科学院院长基金、国家杰出青年科学基金、王宽成基金、留学回国人员择优支持基金和云南省应用基础研究基金等的资助

本文 1995 年 6 月 28 日收到，同年 11 月 3 日修回

频率可能很高。在这种情况下，排序过程变得十分复杂，一般必须借助于计算机。各种主要的 DNA 序列分析软件中，如 PC／GENE,GCG 和 MacVector 等，都有 DNA 排序功能。根据我们的经验，如果 DNA 同源度低于 70%-75%，就不易获得确定的排序。从图 1 中可看出，不同的排序代表了不同的进化途径。从序列 a 到序列 b，最少需 3 步（a—b_1：第 3 位点的 1 次缺失，第 8、9 位点的 1 次双碱基缺失以及第 14 位点的 1 次转换)，而最多则需 6 步（a—b_4：第 2、7 和 8 位点各自 1 次单碱基缺失，第 2、9 和 14 位点各自 1 次转换)。采用不同的排序，可能得到完全不同的系统树。一种稳健的方法是，删除涉及缺失／插入的序列片段。但是，有时缺失／插入可能代表重要的进化信息，简单的删除并不可取。我们建议，如果存在多种合理的排序，而不同的排序又得到不同的系统树，就应该再测定另一个独立的 DNA 片段序列，根据这段序列得到的系统树判断究竟哪种排序更为合理。如果无法获得新的序列，增加外源物种数可能有助于问题的解决。

2 "联合"（Combiend）还是"一致"（Consensus）

在确定 DNA 序列的排序后，我们经常会面临如何处理多组 DNA 数据的问题。为了获得物种树，研究来自不同基因区域的多组序列数据是必不可少的。在获取多组序列数据后，有两种处理方法。第 1 种是先根据各个基因序列，分别构建各自的分子系统树。然后根据这些分子系统树的共同之处，构建"一致"的系统树("一致"途径)，(Peng 等，1982)。以"严格一致"（strict consensus）的系统树为例，该系统树中的每个分枝(branch 必须是在所有单个分子系统树中都完全相同的分枝。换言之，任何在各个分子系统树间有分歧的分枝，都不能为"严格一致"系统树所接受。第 2 种是先把多组序列数据合并为一组，然后在这一合并的序列数据基础上构建系统树（联合途径）(Kluge,1989)。

```
b1  CG-TAGT--CATGAC
b2  CG-TAG-T-CATGAC
a   CGATAGTTCCATGGC
b3  C-GTAGT--CATGAC
b4  C-GTAG--TCATGAC
```

图 1 DNA 同源序列 a 和 b 的排列

Fig.1 The alignment of DNA sequencesa and b

b_1—b_4 代表部分可能的不同排列方式

(b_1,b_2,b_3 and b_4 demonstrate some possible alignments)。

主张"一致"途径的认为，通过该种方法有时能获得比较稳妥的系统树。另外，当某一基因序列数据特别多时，如果采用"联合"途径，该基因的数据可能会掩盖其它基因的进化信息。换言之，在这种情况下，根据多基因系列数据构建的系统树，可能实际上仅代表了某一基因的进化信息。而"一致"途径赋予每个基因片段（不论序列长短）相等的加权值，因而有助于防止这种偏差(Kluge,1983；Hillus,1987)。然而，如果赋予各基因片段相等的加权值，相应地各片段内每一位点的加权值就取决于该片段的相对长短和变异度。片段较长、变异度较高的，其每一位点的加权值就较低。反之亦然。这实际上将导致对不同序列位点的随意加权(Cracarft 等,1989)。同时，"一致"途径并不一定稳妥。图 2 将有助于我们理解"一致"的问题。与正确的物种树（图 2d）比较，A 基因序列揭示了物种 1 和 2 的正确关系，但未能解决它们与物种 3 和 4 之间的关系（图 2a)。类似地，B 基因序列仅揭示出物种 3 与 4 之间的关系而没能解决它们与物种 1 和 2 之间的关系（图 2b)。换言之，基因 A 和 B 的综合已包含了所有 4 个物种间的正确进化关系。然而，基因树 a 和 b 的"严格一致"虽然稳

妥无错，却未能解决这4个物种间的任何关系（图2c）。"联合"途径则有完全不同的特点。由于直接利用所有序列位点提供的进化信息，通过"联合"途径获得的结果可能会更接近于正确的物种树。这有两方面的原因：1）一些物种间的进化关系仅显示于某些基因中。以熊超科为例，现在还没有发现，实际上恐怕也不存在一个理想的基因——仅靠该基因就能揭示所有熊超科物种间的进化关系。换言之，即使对于一个动物超科，也难免需要多基因的"联合"。2）当不同序列位点具有相互抵触的信息时，增加序列位点数有助于显示出正确的进化信息，排除进化杂音（Queinz,1993）。从图2来看，采用"联合"途径分析A基因和B基因序列，就极有可能获得正确的物种数d。在我们对熊超科7个种的研究中，就出现了类似图2中的现象（Zhang等，1994）。

这两种途径选择的关键在于各序列位点是否独立。当各位点严重不独立时，同一基因内的不同位点更有可能倾向于支持某一错误的进化关系。此时，"一致"途径恐怕是合理的选择。如果没有理由怀疑各位点的独立性，"联合"途径是理想的选择。实践中，即使在有少数序列位点不独立的情况下，也应该先考察基因树。如果各基因树间有冲突，且有冲突的分枝置信度较高时，应采用"一致"途径。而有冲突的分枝置信度较低时，可结合使用"一致"和"联合"两种途径。如果各基因树间没有冲突，建议采用"联合"途径（Zhang等，1994）。

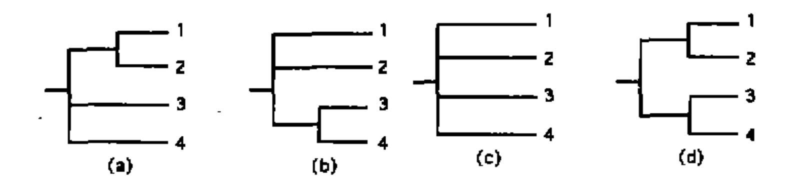


图2　物种1、2、3和4的系统树

Fig. 2　Phylogenetic trees for species 1,2,3 and 4

a. 根据A基因序列构建的分子系统树(based on A gene sequences);
b. 根据B基因序列构建的分子系统树(based on B gene sequences);
c. 系统树a和b的"严格一致"系统树(strict consensus of tree a and b);
d. 正确的物种树(species tree)。

3　DNA序列的加权

在着手构建分子系统树之前，我们还应当了解如何处理不同的DNA位点以及不同的序列变化。现已清楚地知道，为了排除杂音以获取正确的进化信息，在一些情况下对DNA位点及各种序列变化予以加权是必不可少的。以下是一些基本的加权规则。首先，应当区别3种不同类型的序列变化：转换、颠换和缺失／插入。这种区分对于使用具有较高转换／颠换比率的DNA作远缘物种的比较时尤其重要。我们知道，哺乳动物系统重建中最常用的线粒体基因组，其转换／颠换的比率就普遍较高（Zhang等，1993）。一般常用的加权法则是，先确定转换／颠换的比率R，若赋予转换的加权值为1，颠换的加权值则为R。对于远缘物种的比较，如果转换已趋于饱和，就可完全忽略转换而仅使用颠换。至于近缘物种的比较，由于转换一般尚未趋于饱和，因而并不包含很多进化杂音。此时，为了方便起见，根据我们的经验，可以赋予转换和颠换相同的加权值。对缺失／插入的加权还没有较为统一的看法。多碱基的缺失／插入，在没有明显证据表明是由多次缺失／插入积累而形成的时，一般可以认为是由多个碱基的一次缺失／插入形成的。也就是说，可以作为1个变异特征。在缺失／插入很少发生的区域，如蛋白质编码区，缺失／插入的加

权值或许应高于颠换。而在缺失／插入经常发生的区域，其加权值或许近似于转换。也有人主张干脆忽略涉及缺失／插入的区域。其次，我们应当区别不同的 DNA 位点。就蛋白质编码基因而言，密码子第三位点进化最快，应赋予最低的加权值；而密码子第二位点最保守，故应赋予最高的加权值。以哺乳动物线粒体细胞色素 b 基因为例，其常用的加权方法是，密码子第一和第二位点取代（包括转换与颠换）的加权值等同于第三位点颠换的加权值。而第三位点转换的加权值则为零（Lrwin 等,1991)。在同一蛋白质基因内，一般不再进一步细分保守区域与快速进化区域。一方面这种细分过于复杂，操作上十分困难。另一方面，如此细分也许并非必须。对于 RNA 基因，由于功能上的需要必须维持其二级结构的稳定，因此茎区(stem)与环状区(loop)的进化规律有所不同。总体上而言，茎区似乎相对保守一些。在茎区，当一个位点的突变固定后，与其相对应的位点可能倾向于出现相应的变异，以维持 Watosn-Crick 碱基配对的结构。换言之，在茎区当一个位点发生突变并固定后，可能与其对应位点所发生的相应变异就很容易被保留下来。因此，Wheeler 等(1988)建议对茎区的取代赋予较低的加权值。我们对熊超科线粒体 DNA 序列的研究表明，茎区的多数位点并不遵循配对取代规律(Zhang 等,1993)。目前尚不清楚应赋予茎区多低的加权值。Dixon 等(1993)建议采用 0.8 这一经验值。至于多数的非基因区域，由于我们对其进化规律所知甚微，因此，如何加权还有待探索。

4 构建分子系统树的主要方法

解决上述问题之后，即可利用数学方法，综合分析 DNA 序列，提取进化信息。这也就是通常所说的构建分子系统树。随着系统构建方法的发展及其计算机程序化，从最近两年开始，国际上的有关主要刊物已逐渐要求文章对其所采用的计算机程序及分析过程有清楚的交待。从现有的结果看，在一些情况下，根据同一序列数据，采用不同的构建方法，有可能得到相互矛盾的系统树。

主要的系统重建方法可归纳为 3 类（Nei,1987;Felsenstein,1988）：1）简约法（parsimony methods)。其中最有影响的是最大简约法（Fitch,1977)。这类方法旨在确定最短的系统树——该树仅需要最少的进化步骤就能解释所有 DNA 序列间的变异。对于某一 DNA 序列数据，最短的系统树可能只有一个，也可能有多个。这类方法允许缺失一些分类单元的部分 DNA 序列数据。也就是说，即使在无法获取少数分类单元完整的 DNA 序列的情况下，仍可使用简约法。专为简约法设计的计算机软件 PAUP(Swofford,1993)功能很全，操作方便，是目前最佳和最有影响的简约法软件。2）距离法（distance methods)。这类方法首先需要从 DNA 序列计算每对分类单元间的遗传距离。Jukes 等(1969)的单参数法和 Kimura(1980)的双参数法较为常用。软件 PHYLIP(Felsenstein,1993)中的 DNADIST 程序包括了这两种计算距离的方法。在获取距离矩阵后，距离法按照一定的规则，根据各距离值间的内在关系构建系统树。距离法有很多种，其中以 UPGMA 法（Sneath 等,1973）和 Neighbor-joining 法（Saitou 等,1987）影响最大。这两种方法都可用 PHYLIP（Feleenstein,1993）中的 NEIGHBOR 程序进行计算。3）似然法(likelihood methods)。这类方法首先需要确定一个序列进化的模型，如 Kimura(1980)的双参数模型等。然后寻找在该进化模型下，最有可能产生所研究 DNA 序列数据的系统树。这类方法要求所有研究的分类单元都具有完整的 DNA 序列数据，在运算过程中仅考虑碱基取

代而忽略缺失／插入。由于这类方法的计算特别复杂费时，因此其应用并不如前两类方法那么普遍，最大似然法（maximum likelihood method）是其中影响最大的一种(Felsentein,1981)。这种方法可用 PHYLIP（Felsenstein,1993）中的 DNAML 程序进行分析运算。

应当指出的是，通过上述方法获得的分子系统树是无根的（unrooted tree)。但是，我们可以通过外群分析确定树的根。

在用某种方法获取系统树后，还有必要用重抽样法（bootstrap）评估系统树的可靠性（Felsenstein,1985)。这种方法的操作一般都需要计算机。PAUP（Swoffird,1993）和 PHYLIP（Felsentein,1993）两种软件中都有重抽样分析的功能，不过运算较为费时。

值得注意的是，各类方法都需要一定的前提条件，因而也有一定的运用范围。然而，我们对许多条件所知甚微，因而很难判断在某一具体情况下哪种方法最佳(Felsenstein,1988)。我们认为，最好同时合用多类方法构建系统树。多种方法所獲系统树的一致，将大大提高结果的可靠性（Kim,1993;Zhang 等,1994)。

对于上述各类方法的详细原理及其限制，可参阅 Swofford 等(1990)。至于各种计算机软件的功能和具体操作方法，如特征加权，树的搜寻、重抽样分析等，作者拟另外撰文介绍，在此不再赘述。

5 基因树与物种树

当一个分子系统树是根据某一基因数据构建而来时，就称为基因树。物种树则是指代表了一组物种进化过程的系统树(Nei,1987)。基因树与物种树可能存在两方面的区别。1）对于某一被研究的基因，可能存在种内的多态性。换言之，在物种分化之前，该基因可能已开始分化。因此，两物种间该基因的分化时间可能早于这两个物种的分化时间。由这一基因计算而来的分枝长度（分歧时间）可能偏高。对于较长时间的进化过程而言，因种内多态导致的这种误差可以忽略不计。但是，对于新分化的物种、这种误差的影响可能很大。2）基因树的分枝情况（拓扑结构）可能不同于物种树的。这种情况一般发生在分枝点非常接近的物种间。人、猩猩和大猩猩间的关系可能是较为典型的例子。这是因为 DNA 突变是随机过程，在有限的序列内可能存在统计学上的偏差。通过增加 DNA 序列的长度并测定多个相互独立的基因片段，一般可以避免这种问题的发生。

我们所研究的物种进化过程都已成为历史，我们不可能重建出绝对完整的历史，同样也不可能获取绝对的物种树。但是，通过多基因、大量 DNA 序列的正确分析，可以最大限度地缩小基因树与物种树间的差别。在这种情况下获得的系统树一般也可被接受为物种树。

致谢 本所王应祥和刘瑞清教授、本实验室刘爱华教授、王文、宿兵、聂龙、和朱春玲等同志对本工作给予热情的支持与帮助。特此致谢！

参 考 文 献

张亚平，1995．熊超科DNA序列进化及其保护生物学意义．见：中国科学技术协会第二届青年学术年会论文集．中国科学技术出版社．462—467.

Cracaft J, Mindell D P, 1989. The early history of modern birds: a comparison of molecular and morphological

evidence. In: Frenholm B, Bremer K, Jornvall H, eds. The hierarchy of lift. Elsevier, Amsterdam: 389-403.

Dixon M T, Hillis D M, 1993. Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. *Mol. Biol. Evol.*,**10**: 256-267.

Felsenstein J, 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**: 368-376

Felsenstein J, 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**: 783-791.

Felsenstein J, 1988 Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.*, **22**: 521-565.

Felsenstein J, 1993. "Phylip (Phylogeny inference package)" version 3.5c. University of Washington.

Fitch W M, 1977. On the problem of discovering the most pasimonioustree. *Am. Mat.*, **111**: 223-257.

Hillis D M, 1987. Molecular versus morphological approaches to systematics. *Annu. Rev. Ecol. Syst.*, **18**: 23-42.

Hillis D M, Moritz C, 1990. An overview of applications of molecular systematics. In: Hillis D M, Moritz C, eds. Molecular systematics. Sunderland, Massachusetts: Sinauer Associates Inc. 502-515.

Irwin D M, Kocher T D, Wilson A C, 1991. Evolution of the cytochrome b gene of mammals. *J. Mol. Evol.*, **32**: 128-144.

Jukes T H, Cantor C R, 1969. Evolution of protein molecules. In: Munro H N, ed. Mammalian protein metabolism. New York: Academic Press. 21-132.

Kim J, 1993. Improving the accuracy of phylogenetic estimation by combining different methods. *Syst. Biol.*, **42**: 331-340.

Kimura M, 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*,**16**: 111-120.

Kluge A G, 1983. Cladistics and the classification of the great apes. In: Ciochan R L, Corruccini R S, Eds. New interpretations of ape and human ancestry. New York: Plenum. 151-177.

Kluge A G, 1989. A Concer for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, serpentes). *Syst. Zool.*, **38**: 7-25.

Miyamoto M M, Cann L, Allen D *et al*, 1987. Phylogenetic relationships of humans and African apes as ascertained from DNA sequences (7.1 kilobase pairs) of the globin region. *Science*, **238**: 369-373.

Nei M, 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Penny D, Foulds Z R, Henny M D *et al*, 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*, **297**: 197-200.

Queiroz A, 1993. For consensus (sometimes). *Syst. Biol.*, **42**: 368-372.

Saitou N, Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol. Evol.*, **4**: 406-425.

Sneath P H A, Sokal R R, 1973. Numerical taxonomy. San Francisco: Freeman.

Swofford D L, Olsen G J, 1990. Phylogeny reconstruction. In: Hillis D M, Moritz C, eds. Molecular systematics. Sunderland, Massachusetts: Sinauer Associates Inc. 411-501.

Swofford D L, 1993. "PAUP: phylogenetic analysis using parsimony". version 3.1, Illinois Natural History Survey, Champaign.

Wheeler W C, Honeycutt R L, 1988. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. *Mol. Biol. Evol.*, **5**: 90-96.

Zhang Y P, Ryder O A, 1993. Mitochondrial DNA sequence evolution in the Arctoidea. *Proc. Natl. Acad. Sci. USA*, **90**: 9557-9561.

Zhang Y P, Ryder O A, 1994. Phylogeny relationships of bears (the Ursidae) inferred from mitochondrial DNA sequences. *Mol. Phylogenet. Evol.*, **3**: 351-359.